Topic: Clustering
Devising a House Improvement Plan for a Hotel Complex in Mallorca

BIDS, gwendolin.wilke@hslu.ch

Your data refers to a questionnaire among customers of a hotel complex in Mallorca that consists of 30 separate houses. In each of the houses customers were asked to give feedback on several aspects of the house and its service quality. Your data set contains for each house the percentage of positive answers per question[1]. The goal of the hotel management is to improve the customer ratings of the hotel complex. Your task is to explore the data to find out which aspects and which houses are most problematic, so that the management can set priorities.

Data Set Description

Name: questionnaire.csv
Attributes: They refer to the questions in the questionnaire and are measured in % of positive answers.

- *value*. Overall value for money.
- *complaint.* Handling of customer complaints.
- *facilities*. Bathroom.
- *clean.* Cleanliness.
- *athm.* Positive atmosphere.
- *service.* Service quality (reception, maid service, restaurant)
- *amenities.* TV, WLAN, parking, minibar, dining, pool, fitness room.

Assignment

1. Load the data set and view it using the `View()` command.
2. Discuss the dataset based on `str()` and `summary()`. What does it tell you about the houses?[2] I.e., what are the most problematic aspects (attributes) the management should focus on?
3. Apply K-means with 4 clusters.
4. Evaluate your clustering results using the evaluation measures discussed in class. Interpret them.
5. Interpret your clustering results:
   - What is the biggest, what is the smallest cluster?[3]
   - Inspect the coordinates of the `cluster means` (i.e. the cluster centers) and interpret them: What are the groups of houses the management should focus on, and why?[4] For each group of houses, what are the most problematic aspects?
   - Estimate the average percentage of positive answers per cluster (average over the attribute values of each cluster center-point) to get an overall performance KPI for each cluster of houses. Which cluster has the best performance on average?
6. Try some other values for the number of clusters (instead of 4). Write down the resulting `withinss` values and draw the respective graph[5]. What do you observe? Are 4 clusters a good choice?

[1] An answer was counted as 'positive', if the assigned point score exceeded 65% of the maximum possible points.

[2] If you have never heard of Quartiles or the Median, look it up on Wikipedia. For each attribute in the summary, what is the percentage of positive answers for each quartile? What does it tell you about the houses? (E.g., if 75% of all houses (this is the 3rd Qu., i.e. most houses) achieve not more than 47.75% percent of positive answers for amenities, you know that amenities is a problematic aspect for most houses, and the management may want to start their improvement efforts here…)

[3] You'll find the cluster sizes in the first line of your result variable.

[4] Each cluster center is a point in your 7-dimensional attribute space. For each cluster center-point, go through its coordinates. (E.g., if the coordinate value of the cluster center point of cluster 1 are all rather high, you know that the houses in this cluster tend to get a high percentage of positive answers to all questions. These houses need not much or no improvement…)

[5] X-axes: Number of cluster centers, y-axes: corresponding `withinss` values.